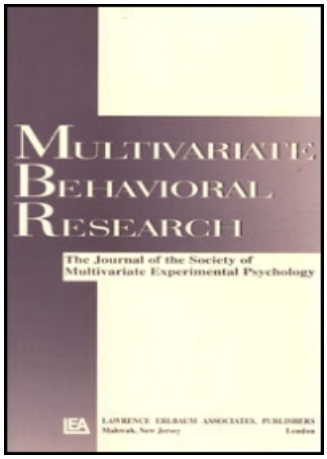
Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse

Bernaards, C.A.; Sijtsma, K.

Published in:
Multivariate Behavioral Research

Publication date:
1999

Document Version
Publisher's PDF, also known as Version of record



Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653673

Factor Analysis of Multidimensional Polytomous Item Response Data Suffering From Ignorable Item Nonresponse

Coen A. Bernaards [a]; Klaas Sijtsma [b]
[a] Department of Methodology and Statistics FSW, Utrecht University, the Netherlands.
[b] Department of Methodology FSW, Tilburg University, the Netherlands.

Online Publication Date: 01 July 1999
To cite this Article: Bernaards, Coen A. and Sijtsma, Klaas (1999) 'Factor Analysis of Multidimensional Polytomous Item Response Data Suffering From Ignorable Item Nonresponse', Multivariate Behavioral Research, 34:3, 277 - 313
To link to this article: DOI: 10.1207/S15327906MBR3403_1
URL: http://dx.doi.org/10.1207/S15327906MBR3403_1

Multivariate Behavioral Research, 34 (3), 277-313

Factor Analysis of Multidimensional Polytomous Item Response Data Suffering From Ignorable Item Nonresponse

Coen A. Bernaards
Department of Methodology and Statistics FSW, Utrecht University, the Netherlands

Klaas Sijtsma
Department of Methodology FSW, Tilburg University, the Netherlands

This study deals with the problem of missing item responses in tests and questionnaires when factor analysis is used to study the structure of the items. Multidimensional rating scale data were simulated, and item scores were deleted under Rubin's (1976) MAR and MCAR definitions. Five imputation methods, the *EM* algorithm, and listwise deletion were implemented to deal with the item score missingness. Factor analysis was done on the complete data matrix, and on the seven data matrices that resulted from the application of each of the missingness methods. The factor loadings structure based on *EM* best approximated the loadings structure obtained from the complete data. Imputation of the mean per person across the available scores for that person was the best alternative to *EM*. It is recommended to researchers to use this simple method when *EM* is not available or when expertise to implement *EM* is lacking.

Introduction

Factor analysis is often used to study the structure of the item set in tests and questionnaires. A well known and difficult problem in data collection via tests and questionnaires is item nonresponse. Item nonresponse occurs if respondents are unable or reluctant to provide answers to one or more items or if they accidentally skip items, but at the same time produce answers to other items. In this article, we are concerned with item nonresponse which can be considered to be a completely random phenomenon in the population at hand (responses are missing completely at random, MCAR; Little & Rubin, 1987, pp. 14-15), or which is a random phenomenon in particular well-

The authors would like to thank Computer Algebra Nederland (CAN) computer facilities, Amsterdam, for providing access to *S*-Plus.

Correspondence to: Coen A. Bernaards, Utrecht University, Faculty of Social Sciences, Department of Methodology and Statistics, P.O. Box 80140, 3508 TC Utrecht, The Netherlands; tel: +31-30-2539138; fax: +31-30-2535797; e-mail: C.A.Bernaards@fss.uu.nl

defined subgroups of the population but which may vary in degree across such subgroups (responses are missing at random, MAR; Little & Rubin, 1987, pp. 14-15). An example of a completely random phenomenon is that respondents accidentally skip questions. An example of random item nonresponse in subgroups is that older respondents tend to accidentally skip more questions than younger respondents. Here, age is a covariate that explains differences between meaningful subgroups. The response mechanisms MCAR and MAR are known to produce *ignorable* nonresponse.

Other relevant forms of item nonresponse exist, but are postponed to a later study (Bernaards & Sijtsma, 1999). For example, a respondent may be unable to give an answer when he or she lacks adequate information, and refuses to guess or otherwise give a fake answer. If this tendency is typical of this respondent and not of others, this kind of item nonresponse does not fall under the definitions of MCAR and MAR. Other examples of nonresponse that do not subsume under these definitions are reluctance to respond if, for example, a question is considered menacing to privacy (questions about ones sexual habits or income) or embarrassing (questions about the relationship with one's parents or children) when these opinions are not typical of the whole population or of particular subgroups. Item nonresponse does not include refusal of respondents to take part in the investigation, known as unit nonresponse, or dropout from the investigation due to illness, moving to another city, and so on, known as experimental mortality.

Thus, we consider the case when all respondents produced answers to at least some of the items, but not all respondents gave answers to all items. If nothing is done about item nonresponse, this may highly influence results from factor analysis and other multivariate statistical analyses, since incomplete cases will simply be omitted from the data to prevent covariance matrices from not being positive (semi)definite. Often the causes of item nonresponse are unknown to the researcher. If item nonresponse is not a random process, after omission of incomplete cases the reduced data matrix may no longer be representative of the population of interest. Therefore, it is important to deal with item nonresponse in a sensible way, for example, by estimating the missing item scores. Other reasons may be that a larger sample leads to more accurate estimates of parameters and an increased power for testing hypotheses.

In this study, we investigated missing data problems in the context of factor analysis of questionnaire rating scale data. Other researchers have addressed missing data problems with factor analysis. Cattell (1978, pp. 515-516) discussed six ways of dealing with missing item scores in

questionnaires, among which were listwise deletion, item mean imputation, and imputation based on multiple regression. Finkbeiner (1979) investigated five methods, for example, listwise deletion, an item mean replacement method, and a principal components method (also see Huisman & Molenaar, 1997). Brown (1983) compared listwise deletion, pairwise deletion, regression estimation, and the maximum likelihood method for estimating the loadings from a single factor model. For MCAR, Lee (1986) compared listwise deletion, generalized least squares estimation, and maximum likelihood estimation in structural equation models. Also in a structural equation modeling context, Muthén, Kaplan, and Hollis (1987) compared several methods for dealing with missingness that is not completely random. Rather than imputing values for missing item scores or handling the missing data problem in the likelihood function, one could under certain assumptions estimate the complete data correlation matrix of the items, and then perform factor analysis directly on this estimated matrix; for example, see Wilks (1932), Timm (1970), Gleason and Staelin (1975), and Knol and Ten Berge (1989) for more information.

In this article, we study the performance of five imputation methods designed to deal with missing item scores in questionnaire data, and two missing data methods not based on imputation (listwise deletion and the *EM* algorithm). For simulated questionnaire data containing missings, the question was how well the use of these methods to produce complete data can lead to the reconstruction of the factors that resulted from the original complete data. Correspondence between loadings matrices based on complete data and loadings matrices based on data also containing imputed scores was evaluated using several indices. This led to recommendations concerning the use of missing data methods in practical questionnaire research where factor analysis of the data is envisaged.

The simulated data were the scores on a test or questionnaire consisting of ordered five-point rating scales (Likert items) and, for each person, scores on two covariates. Although many test theory models assume that all items measure the same trait (unidimensionality), in practice responses frequently are the result of a combination of latent traits (multidimensionality). For example, responses to an item on introversion could partly be determined by language skills. Most simulated item score data matrices used here were multidimensional, while a smaller number were unidimensional. Complete simulated data matrices were generated by means of a multidimensional polytomous item response theory (IRT) model (Kelderman & Rijkes, 1994) and, next, subjected to factor analysis. Dolan (1994) demonstrated that even with five-point rating scale data, factor analysis is not seriously affected by deviations from normality of the distributions of the variables. Takane and

De Leeuw (1987), Muraki and Carlson (1995), and McDonald (1997) discussed the relation between multidimensional IRT and factor analysis; see Bello (1993) and Knol and Berger (1991) for related work in this area.

Next, item scores were deleted both under the MCAR and MAR definitions of missingness (Little & Rubin, 1987; to be discussed later on in full detail), and the resulting incomplete data matrix was then treated by subsequently applying one of the seven missing data methods, which yielded seven reconstructed data matrices. For each of these seven data matrices the same number of factors was extracted as for the complete data matrix. This way, factor analysis results based on a complete data matrix could be compared with the results obtained under each of the missing data methods, and conclusions could be drawn on the effectiveness of these methods in producing the correct results.

Method

Generating the Data

The multidimensional polytomous latent trait (MPLT) model (Kelderman & Rijkes, 1994) was used to generate the polytomous item scores. A two-step procedure was followed. First, for each simulee latent trait values were drawn from a multivariate distribution (multidimensionality). This determined the number of factors underlying the data. Second, given these latent trait values the MPLT model was used to determine for each combination of a simulee (defined by a combination of latent trait values) and an item each of the probabilities of responding in particular answer categories. These probabilities were used to generate the final data. Thus, this two-step procedure generated a multivariate distribution of item scores (step 2) based on a multidimensional latent trait structure (step 1). Also note that the underlying trait structure may, as a special case, be unidimensional. We generated both unidimensional and multidimensional multivariate item score distributions.

Suppose $i = 1, ..., N$ respondents answer to $j = 1, ..., k$ items. Each item has ordered answer categories with scores $x = 0, ..., r$; here, $r = 4$. The items measure a combination of latent traits according to some a priori known ratio per item. For example, ten items may measure latent trait A and latent trait B with weights 1 and 3, respectively, and the next ten items may measure these traits with weights 3 and 1, respectively. Latent traits are denoted by θ with indices i for identifying persons and indices q ($q = 1, ..., s$) for identifying traits, so that the notation is θ_{iq}.

The scoring weights associated with the response categories are contained in the three-way array **B** with entries B_{jqx}. The scoring weights reflect the ratio by which item j measures the latent traits, and also can be interpreted as discrimination indices. Following Kelderman and Rijkes (1994), we maintain the terminology of scoring weights. The separation parameters for the categories associated with B_{jqx} are contained in the array $\mathbf{\Psi}$ with elements Ψ_{jqx}. By choosing the scoring weights **B** appropriately, different models can be defined.

The MPLT model is of the form

$$(1) \qquad P(X_{ij}=x|\theta_{i1},...,\theta_{is})=\frac{\exp\left[\sum_{q=1}^{s}\left(\theta_{iq}-\Psi_{jqx}\right)B_{jqx}\right]}{\sum_{y=0}^{r}\left\{\exp\left[\sum_{q=1}^{s}\left(\theta_{iq}-\Psi_{jqy}\right)B_{jqy}\right]\right\}}.$$

The MPLT model requires that if $B_{jqy} = 0$ then $\Psi_{jqy} = 0$ to ensure uniqueness of the parameters.

The generation of the data used two binary covariates with scores for each simulated person. In practical research, examples could be respondents' gender and membership of majority or minority groups. The relative occurrence of all four possible combinations of scores on the covariates in the population was known. Combinations of scores on covariates are indexed by g, with $g = 1, ..., 4$. When simulating data, for each of the N persons a combination of covariate scores was drawn with probability equal to the relative frequency in the population.

We assumed that different covariate classes are characterized by different means on the θs. This was formulated as follows. Traits were assumed to be distributed according to a s-variate normal distribution (in this simulation study, $s = 1, 2, 4$) with given matrix of dispersion, and means depending on the covariates via

$$(2) \qquad \boldsymbol{\mu}_{\theta_g}=\begin{pmatrix}\mu_{\theta_{g1}}\\ \vdots \\ \mu_{\theta_{gs}}\end{pmatrix}=\begin{pmatrix}c_{11}\times Z_{g1}+c_{12}\times Z_{g2}\\ \vdots \\ c_{s1}\times Z_{g1}+c_{s2}\times Z_{g2}\end{pmatrix}=\mathbf{C}\mathbf{Z}_g,$$

where $\mathbf{Z}_g = (Z_{g1}, Z_{g2})$ is a vector with the binary scores of covariate class g, and **C** is a $s \times 2$ matrix of weights. Different mean vectors $\boldsymbol{\mu}_{\theta_g}$ can be generated by choosing different sets of weights **C**. In this study (to be discussed later on), one choice was maintained throughout.

Since the scoring weight array **B** in model 1 and the array of separation paramaters $\boldsymbol{\Psi}$ were specified a priori, for each item the probabilities of response in all answer categories could be calculated for each vector $\boldsymbol{\theta}$ drawn from the *s*-variate normal distribution and subjected to the transformation in Equation 2. Next, given a vector $\boldsymbol{\theta}$ for each of the *k* items an outcome was drawn from a multinomial distribution with response probabilities (Equation 1) as calculated for the answer categories, resulting in *k* item responses for each of *N* persons.

"Generating" Missing Item Scores

We considered two cases of missingness. First, scores can be considered MCAR (Little & Rubin, 1987, pp. 14-15) when the missing item responses are a random subsample from all observed item scores, with the covariate structure being ignored. Second, scores can be considered MAR (Little & Rubin, 1987, pp. 14-15), when the missing item responses are not a random subsample of all observed item scores, but a random subsample of item scores within classes defined by the covariates. See Greenless, Reece and Zieschang (1982) for a discussion of deviations from the MAR assumption.

In order to compare results from factor analysis of data sets based on different imputation methods and other missing data methods (listwise deletion and *EM*), simulated data matrices each should have the same number of missing item responses. For example, consider a data matrix for 100 respondents who answered 20 items, thus yielding 2000 scores in total. When 20 percent of the item scores out of 2000 item scores are deleted, 400 scores should be missing in each replication.

Under the MCAR assumption, this was realized by randomly assigning, with equal probability, 400 of the 2000 scores in the data matrix the status of missing value. Each imputation method thus replaced each of these 400 missing item scores by an imputed score; listwise deletion omitted the data lines that contained missing item scores; and the *EM* algorithm maximized the likelihood by iterating between updating the missing values given the factor scores and alternatingly estimating of the covariance matrix given the data (*E* step), and estimating the factor loadings and factor scores given the covariance matrix (*M* step).

Under the MAR assumption, covariate classes should be represented among the missing scores in an a priori specified ratio. This was realized by treating the sampling of 400 scores designated as missing as drawing from a multinomial distribution with 2000 categories (corresponding to the entries of the data matrix). Since a person belongs to one covariate class, the

Table 1
Relative Expected Frequency of Nonresponse for Two Binary Covariates, and Corresponding Probabilities

		Covariate 2				Covariate 2	
		0	1			0	1
Covariate 1	0	8	4	Covariate 1	0	.47	.23
	1	3	2		1	.18	.12

separate scores in a data line each had the same probability of being sampled. Table 1 contains an example for two covariates where a respondent with two 0 scores has a probability of not responding to an item that is four times the probability of a person with two 1 scores, and so on.

For this example, the multinomial distribution with 2000 categories has four different probabilities. The desired ratio of these probabilities was fed to a program and the relative frequencies were then normalized to probabilities summing to 1 (see Table 1). This procedure also resulted in the assigning of 400 scores as missing. Note that due to randomly drawing from a multinomial distribution, in samples the ratios of covariate classes as represented among the missing scores may deviate from the theoretical a priori ratio.

Implemented Missing Data Methods

Five imputation methods were used. Imputation methods estimate the unobserved values and then replace the missing value by this estimate. The result is a complete data matrix, and the standard factor analysis estimation procedures can be carried out on this data matrix without loss of cases. We chose imputation methods that were easy to understand and to implement by practical researchers without a training in applied statistics or psychometrics. One of the methods imputed random scores and was used as a benchmark. Of the other four methods, to be defined shortly, Cattell (1978), Finkbeiner (1979), and Huisman (1998) discussed item mean imputation, and in the context of scale construction Huisman (1998) also studied person mean imputation. Moreover, we studied two methods using the total mean and a mean conditional on covariate classes, respectively.

1. *Random Imputation* (*RI*) draws at random from a multinomial distribution with outcomes 0,...,4, and equal probabilities (0.2) for each outcome.

2. *Overall Mean Imputation* (*OM*) calculates the mean across all available item scores in the data matrix and imputes this value for all missings.

3. *Mean Conditional on the Covariates* (*CM*) imputes the mean based on the available scores across all items of all persons within the same covariate class, and imputes this mean for each missing in this covariate class.

4. *Item Mean* (*IM*) calculates for each item its mean across the available scores, and imputes this mean for each missing value of this item.

5. *Person Mean* (*PM*) calculates for each person his/her mean over the available item scores, and imputes this mean for each missing value of that particular person.

Two other methods, not based on imputation, were also implemented because they are widely used in missing data problems.

1. *Listwise Deletion* (*LD*) is popular in the social sciences. For example, *LD* is often used by researchers who analyze their multivariate data by means of the sofware package SPSS. *LD* results in factor analysis based only on those persons who responded to all items.

2. *The EM algorithm* (*EM*) as described by Little and Rubin (1987, pp. 148-149) is widely used among statisticians and psychometricians as a method to estimate a covariance matrix subject to missing data.

RI serves as a benchmark for the other methods. If factor analysis results based on another method hardly are better than those based on *RI*, that method should not be used at all.

OM is easy to calculate. However, due to lack of any kind of separation between groups based on covariates, and given that *OM* ignores multidimensionality, its performance may not be much better than *RI*.

Because the means of the latent traits differ across covariate classes, *CM* may yield sensible factor analysis results. However, *CM* is computed ignoring the latent trait structure underlying the data, and this may impair factor analysis results.

Imputation of the mean per item, *IM*, is not impaired by multidimensionality because it is based on single items. Unlike method *CM*, however, *IM* ignores the covariate structure.

PM takes the mean over the smallest meaningfull group of item responses of all imputation methods discussed here and, therefore, it may be the least biased but it may have the largest variance. However, in case of multidimensionality bias also may be large, especially if the correlation between latent traits is small. In case of unidimensionality this method may be expected to perform well.

LD can reduce the number of valid respondents dramatically. For example, consider a respondent who produced a response to each of 20 items with probability 0.9. Then the probability that he or she answered to all 20 items is $0.9^{20} \approx 0.122$. Hence the probability of not responding to at least one item, and thus of being eliminated by *LD*, is 0.878. Another method which is popular in the social sciences is Pairwise Deletion. This method was not implemented because it may result in covariance matrices that are not positive (semi) definite, and because in other situations *LD* is superior over Pairwise Deletion (Kim & Curry, 1977).

Finally, the *EM* algorithm (Dempster, Laird, & Rubin 1977; Little & Rubin, 1987) handles the factor scores from factor analysis as missing. Initially, random values are substituted for the missing data. In the *E* step, the missing values are updated given the factor scores, and the expected value of the covariance matrix given the factor loadings is calculated. In the *M* step the factor loadings and factor scores are updated based on the current estimate of the covariance matrix. These two steps are re-iterated until convergence of the likelihood occurs. The *EM* implementation used here is described in detail in the Appendix.

Calculation of a mean usually will not generate an integer. The imputed mean values thus are not "valid" in this sense. However, the present study was concerned only with results from factor analysis and not with the imputed values themselves. Moreover, rounding of imputed values to the nearest integer would probably introduce more error in the data.

Other methods, such as multiple imputation (Rubin, 1987) and computerintensive methods as described by Tanner (1996), are more difficult to understand and to implement and, therefore, were not considered here. Acock (1997) gives an elementary introduction to missing data methods used with social science data. Little and Rubin (1987) provide an elaborate treatment of missing data methods. Other sources on missing data methods are, for example, Little and Rubin (1989), Rubin (1991), Little and Schenker (1995), Rubin (1996), and Schafer (1997).

Performance of Imputation Methods

One method for assessing performance of an imputation method in factor analysis is the coefficient of congruence or Tucker's ϕ (Tucker, 1951; Ten Berge, 1977), defined as

$$(3) \qquad \phi(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T\mathbf{b}(\mathbf{a}^T\mathbf{a}\mathbf{b}^T\mathbf{b})^{-\frac{1}{2}}, \qquad \mathbf{a}, \mathbf{b} \in \mathbb{R}^k \qquad (\mathbf{a} \neq 0, \mathbf{b} \neq 0).$$

Here **a** and **b** are column vectors containing the loadings of the same *k* variables on one factor obtained in two different data sets, for example, a complete data set and a corresponding data set with imputed values for the missing scores. $\phi(\mathbf{a},\mathbf{b}) = 1$ if and only if **a** is a multiple of **b**. Tucker's ϕ yields a value between the bounds of -1 and 1 for each pair of corresponding factors. For practical purposes, factors with values of ϕ higher than 0.85 are considered to be equal (Ten Berge, 1977; Niesing, 1997). This value serves as a guideline more than an absolute standard.

Another method is defined as follows. Let D^2 denote the sum of squared differences, divided by the number (m) of factors extracted, between the factor loadings on all extracted factors based on the complete data set and the corresponding factor loadings based on the missing data method. The smaller D^2, the more similar the two loadings matrices are; they are exactly identical if $D^2 = 0$. Although there is a positive maximum to D^2 determined by the bounds of the loadings, it is difficult to say when a value of D^2 should be interpreted as high. In our simulation studies, however, we compared D^2 values and, thus, we were interested only in relative values of D^2. D^2 results in one outcome whereas Tucker's ϕ yields separate outcomes for each retained pair of corresponding factors. From loadings matrices **X** and **Y**, D^2 can be calculated through

$$(4) \qquad D^2 = tr\left[(\mathbf{X}-\mathbf{Y})^T(\mathbf{X}-\mathbf{Y})\right]/m = tr\left[(\mathbf{X}-\mathbf{Y})(\mathbf{X}-\mathbf{Y})^T\right]/m.$$

It may be noted that if $D^2 = 0$ then each $\phi = 1$; however, the reverse need not be true. For positive values of D^2 and values of the ϕs lower than 1, there is no exact relation between both indices.

Both Tucker's ϕ and D^2 share the following problem. If, for example, two factors are retained, the result from factor analysis of the complete data could be such that the first ten items heavily load on the first factor and the next ten items heavily load on the second factor. However, due to sampling fluctuation the factor analysis of the "missing data matrix" does not necessarily yield the same pattern: the first ten items may heavily load on the second factor and the next ten items may heavily load on the first factor. When this interchanging of vectors of loadings occurs, the ϕs will be close to 0 and D^2 will be much higher than 0. In order to solve this problem, Tucker's ϕ and D^2 were calculated for two cases: (a) for the loadings matrices produced by the software, and (b) for the loadings matrices with the columns of the second matrix interchanged. Hence, for D^2 the minimum of the two cases was taken, and for Tucker's ϕ the maximum of the sum of the two ϕs was taken to be the criterion of interest. When four factors were extracted, all possible 4! = 24 permutations of loadings were used to calculate both

D^2 and Tucker's ϕ. For D^2 the permutation with the smallest D^2 was chosen. For ϕ the permutation with the maximum of the sum of the four ϕs was taken.

Design of Simulation Study

The choices of design factors and their specific levels were determined by their importance for evaluating the effectiveness of different methods for producing a valid factor analysis solution. Also, the computer time needed was of interest. Table 2 gives an overview of candidate factors. The upper panel of Table 2 contains the factors that were varied, and the lower panel contains the factors that were held constant. The total number of candidate

Table 2
Factors and Levels Relevant to the Simulation study

Factor	Values
Scoring weights **B**	3 configurations, see Table 3
Sample size	100, 500
Percent missingness	5, 10, 20
Relative expected frequency of nonresponse	2 Configurations, see Table 4
Missing Data Methods	Random Imputation, Overall Mean, Conditional Mean, Item Mean, Person Mean, Listwise Deletion, *EM* algorithm.
Extraction method	Principal component, Maximum likelihood.
Method of rotation	Varimax, Procrustes
Number of latent traits	2
Number of items	20
Number of answer categories	5
Distribution of latent traits, $\boldsymbol{\theta}$	$\boldsymbol{\Sigma}$, see Equation 5 Weights **C**, see Equation 6
Separation parameters $\boldsymbol{\Psi}$	fixed per item

design factors was rather high and using them all to build one comprehensive design would render this design too large. For example, if each of the eleven factors in Table 2 would have two different levels, the design would have $2^{11} = 2048$ entries. Due to such practical limitations, we decided to analyze one larger design and two smaller designs. The large design was intended to be comprehensive, and the smaller designs addressed interesting special cases not included in the large design.

Comprehensive Design

Throughout the comprehensive design, two latent traits were used because this is the simplest multidimensional case. Across the design, the number of items was fixed at twenty, divided into two sets of ten items each. Each item had five ordered answer categories, with scores 0,...,4, comparable to Likert (1932) rating scales. A questionnaire consisting of twenty Likert rating scales was considered to be representative of many questionnaires used in practice. We assumed that the first ten items measured both latent traits in another mixture than the second ten items. These mixtures were manipulated through the choice of the *B* weights from the MPLT: the higher *B* for a particular latent trait, the stronger the relation of the item to that trait. The scoring weights are contained in the three-way array **B**, see Equation 1; three different choices appear in Table 3. For each questionnaire, this table gives configurations (first column) of the scoring weights **B** (fourth column) for the two subsets of items (third column). Note that for each item there are five *B*s corresponding to the five answer categories, respectively.

Figures 1, 2 and 3 (following pages) provide the conditional probabilities of an answer in a response category for the different sets of scoring weights **B**. The configuration Mix 1:0 has ten items which exclusively measure θ_1 and ten items which exclusively measure θ_2 (also see Table 3). The interpretation of Mix 3:1 is that half of the items load three times heavier on θ_1 than on θ_2, and for the other half of the items the situation is reversed. Finally, under Mix 1:1 items depend on θ_1 and θ_2 with the same weights. In fact, this is unidimensionality. Mix 1:0 thus contains "pure" items for each trait but the whole questionnaire measures two traits, and Mix 1:1 contains equally weighted mixtures for all items, that is, unidimensionality applies. The choice of Mix 3:1 reflects the idea that, in practice, items are not pure and questionnaires are often not unidimensional, but rather that items measure some mixture of one dominant trait and one or more "nuisance" traits, resulting in at least two factors. We also tried some analyses for Mix 6:1, and concluded that the results already much resembled results for Mix 1:0; thus, Mix 6:1 already came close to the ideal of pure items. Likewise, Mix 2:1 may be too

Table 3
Scoring Weights **B** for MPLT Model.

Mix	Latent trait	item numbers	**B**
1:0	θ_1	1, ..., 10	1, 2, 3, 4, 5
	θ_2	1, ..., 10	0, 0, 0, 0, 0
	θ_1	11, ..., 20	0, 0, 0, 0, 0
	θ_2	11, ..., 20	1, 2, 3, 4, 5
3:1	θ_1	1, ..., 10	1, 2, 3, 4, 5
	θ_2	1, ..., 10	3, 6, 9, 12, 15
	θ_1	11, ..., 20	3, 6, 9, 12, 15
	θ_2	11, ..., 20	1, 2, 3, 4, 5
1:1	θ_1	1, ..., 20	0, 1, 2, 3, 4
	θ_2	1, ..., 20	0, 1, 2, 3, 4

close to Mix 1:1, in fact reflecting a nuisance trait which becomes too strong relative to the dominant trait. Thus, we decided that Mix 3:1 may be an appropriate case between the extremes of Mix 1:0 and Mix 1:1.

Hence, in performing factor analyses the first two factors were extracted for Mix 1:0 and Mix 3:1, and one factor was extracted for Mix 1:1. It may be noted that we assumed the number of latent traits to be known; in another specialized design to be discussed later on, we investigated whether inspection of the eigenvalues of the correlation matrix would have led to the extraction of additional factors.

The distribution of the two latent traits is decisive for the distribution of the item scores over the answer categories per item. Moreover, uncorrelated traits are unrealistic, and correlations higher than, say, 0.7, seem to be rare. The latent traits thus were assumed to originate from a bivariate normal distribution with matrix of dispersion

$$\begin{pmatrix} 2.5 & 0.6 \\ 0.6 & 2.5 \end{pmatrix}, \tag{5}$$

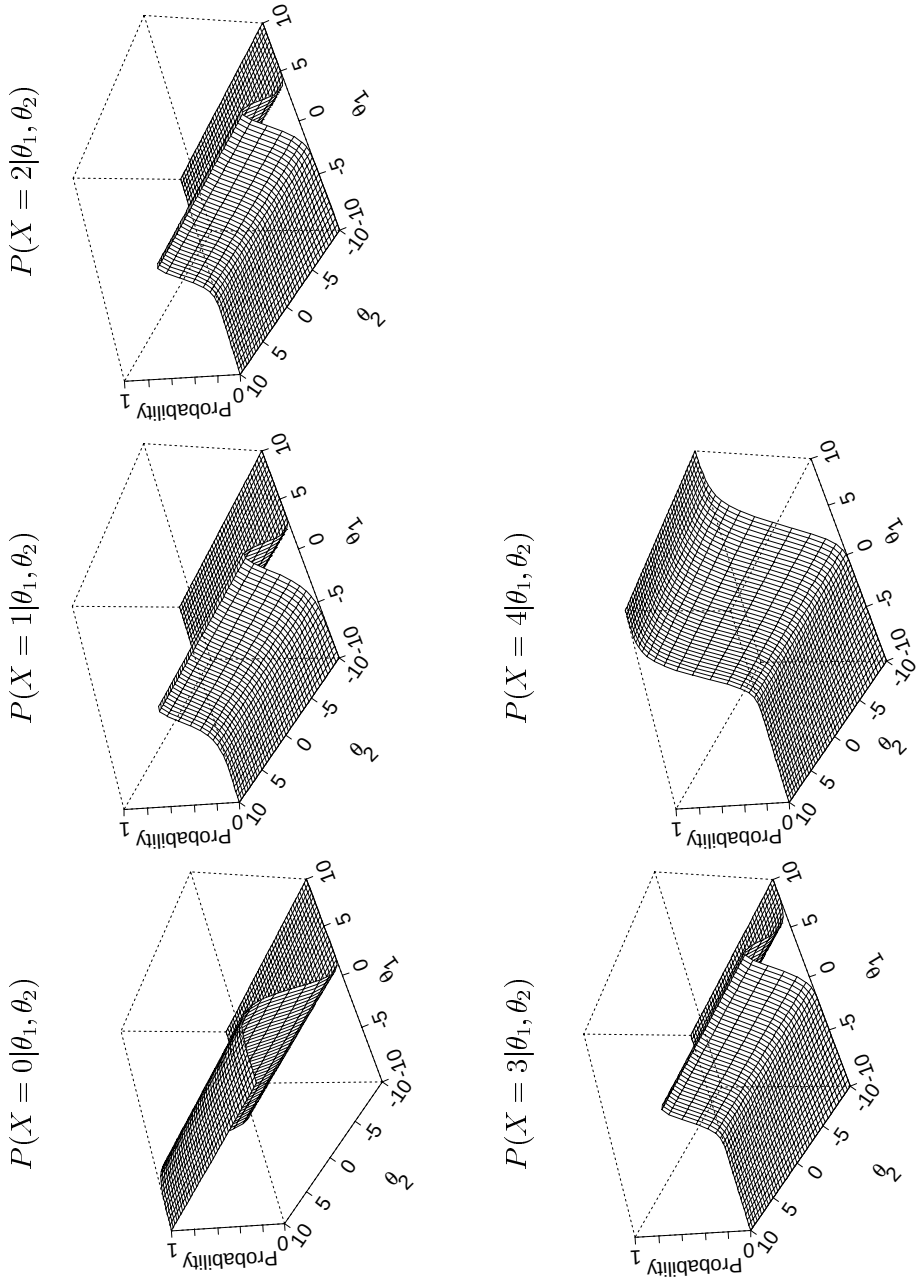


Figure 1
MPLT Model with Scoring Weights Mix 1:0; Consecutive Graphs Give Probability of Score 0, 1, 2, 3, 4, as Function of Two Latent Traits.

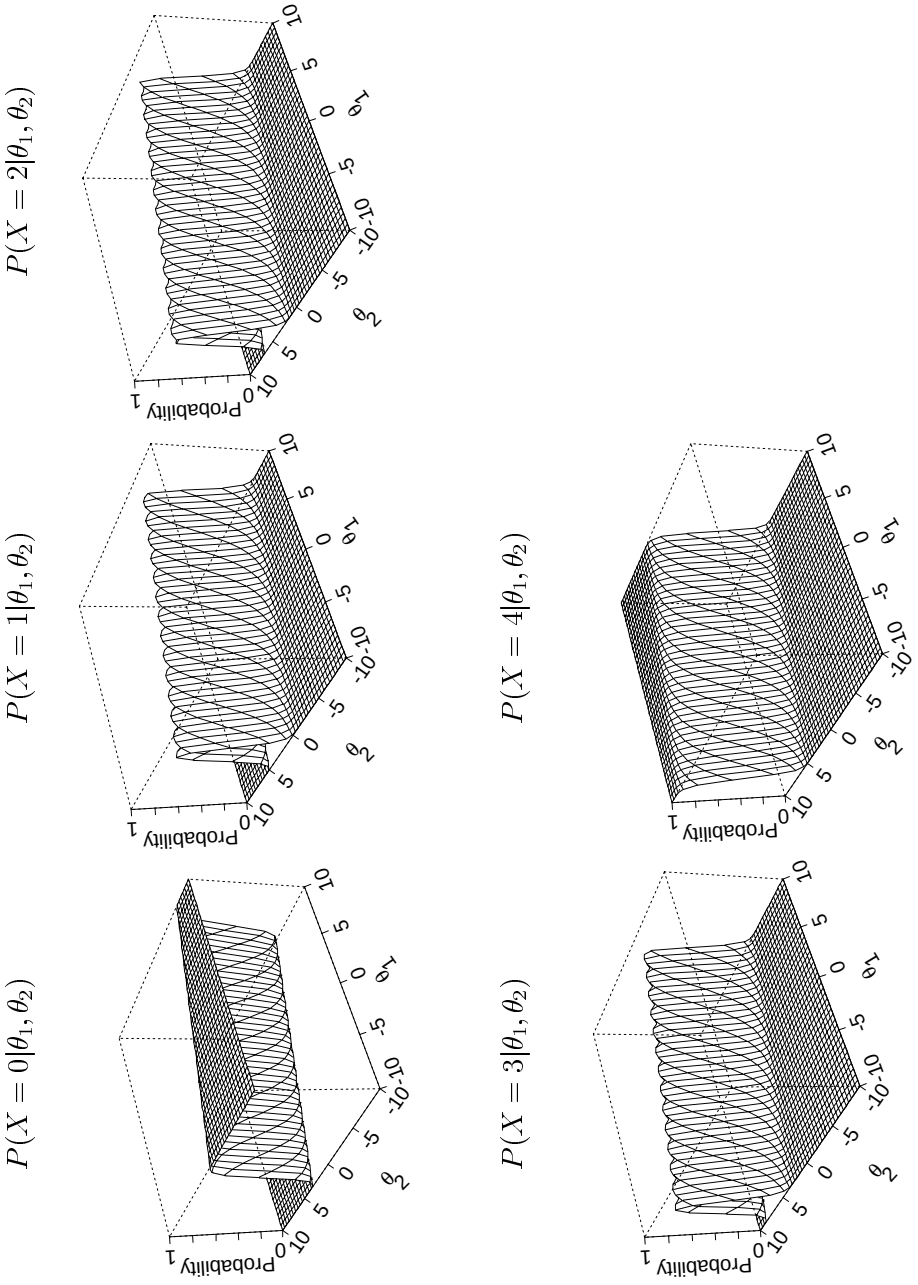


Figure 2
MPLT Model with Scoring Weights Mix 3:1; Consecutive Graphs Give Probability of Score 0, 1, 2, 3, 4, as Function of Two Latent Traits.

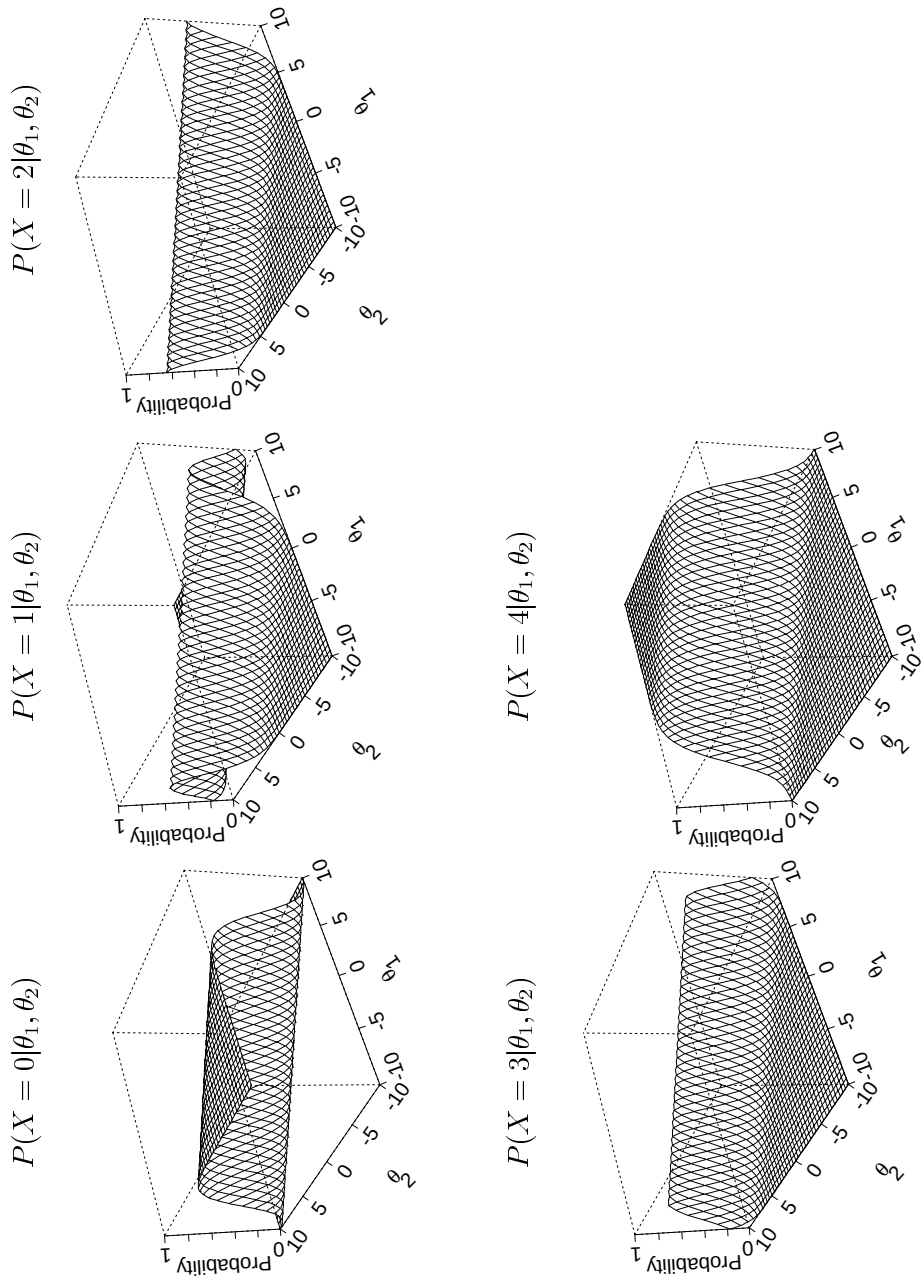


Figure 3
MPLT Model with Scoring Weights Mix 1:1; Consecutive Graphs Give Probability of Score 0, 1, 2, 3, 4, as Function of Two Latent Traits.

from which it follows that the correlation between θ_1 and θ_2 was 0.24. For each covariate class, the choice of matrix **C** implied the means of θ_1 and θ_2; see Equation 2. This matrix was fixed and had values

$$\mathbf{C} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{6}$$

Using Equation 2, this resulted in latent trait means (0,0) for covariate class (0,0); (1,1) for covariate classes (1,0) and (0,1); and (2,2) for covariate class (1,1).

The separation parameters of the MPLT model (Equation 1) were held fixed for each set of scoring weights Mix 1:0, Mix 3:1, and Mix 1:1. This was done in such a way that of the first ten items, on average the first three items had the highest scores (mode = 4), the next four had medium scores (mode = 3), and the last three had the lowest scores (mode = 2). For the next ten items having the reversed weights ratio this was done similarly. More specifically, the item score distributions were realized by choosing the separation parameters to be negative for the items that were to have the highest scores, so that most simulees had high probabilities of producing high item scores; and so on. Figure 4 (next page) gives per panel the histograms for the combined scores on all 6 easiest items, and so on.

To prevent running into computional problems, the sample size should far exceed the number of items. Sample sizes of 100 (small) and 500 (large) were considered to be representative of most questionnaire research conducted in an academic context that uses factor analysis. Also see Dolan (1994) who used simulated data matrices with sample sizes ranging from 200 to 400.

The generation of the missing item scores had two parameters: the total percentage of missingness, and the relative expected frequency of missingness for each set of covariate scores. The present research used 5, 10 and 20 percent missing scores. Application of missing data methods for higher percentages of missing scores was considered undesirable. The missing item scores were generated according to the relative expected frequency of nonresponse for each combination of covariates. Table 4 contains the two configurations used in the simulation study. Configuration REF-MCAR reflects the missing completely at random case; and configuration REF-MAR reflects the missing at random case.

Seven missing data methods were implemented (discussion in the section entitled "Implemented Missing Data Methods").

In psychology, the method of extraction of factors often is principal components analysis. From a mathematical point of view, maximum

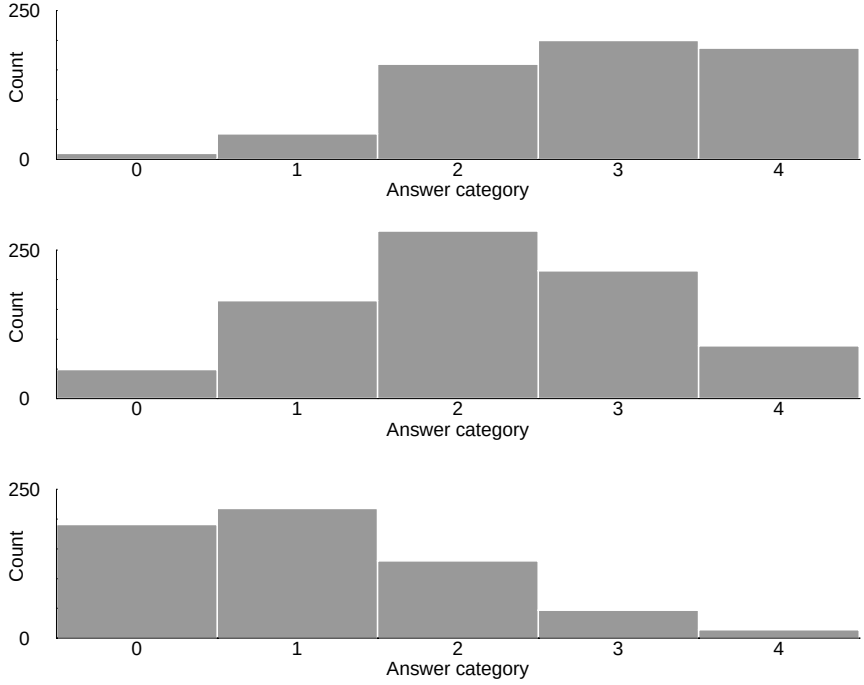


Figure 4
Histograms of Score Distributions of the Total Number of Scores Given by 100 Persons on the 6 Easiest Items (Upper Panel); the 8 Medium Items (Middle Panel); and the 6 Most Difficult Items (Lower Panel).

Table 4
Relative Expected Frequency (REF) of Nonresponse for Combinations of Covariate Scores.

		Covariate 2				
		REF-MCAR			REF-MAR	
		0	1		0	1
Covariate 1	0	1	1	0	5	2
	1	1	1	1	2	1

likelihood estimation of parameters may be preferred. Furthermore, the *EM* algorithm produces maximum likelihood estimates. Hence, both maximum likelihood extraction and principal components extraction were used.

For two-dimensionality cases (Mix 1:0 and Mix 3:1), the first two factors were rotated according to the varimax criterion in order to render them comparable; see for example, Anderson (1984). This seems to be well in agreement with practical factor analysis, where researchers usually rotate their factor solutions to simple structure or another configuration in order to enhance interpretation. In the hypothetical situation where the population loadings structure were known, it would be better to rotate a sample loadings matrix to this population structure. Although not completely comparable, we considered our complete data loadings matrix as the "population" matrix, and for a limited number of design cells also rotated each of the imputed data loadings matrices to this matrix using orthogonal procrustes rotation (Krzanowski, 1988, 159-160). It may be noted, that if $\phi > 0.85$ the results from varimax rotation of the factors based on the complete data and the factors based on the imputed data are comparable (Ten Berge, 1977; Niesing, 1997); thus, in that case procrustes rotation is superfluous. For the one-dimensionality case (Mix 1:1), rotation was not an issue, and D^2 and ϕ were calculated between the factor obtained for the complete data and the factor obtained for the imputed data.

Finally, the results in each design cell were based on 50 replications. The design had 6 factors which were completely crossed, with magnitude 3(scoring weights) × 3(percent missing) × 7(missing data method) × 2(relative expected frequency of nonresponse) × 2(method of extraction) × 2(sample size). This yielded 504 cells.

Specialized Designs

Two specialized designs were analyzed. The first specialized design addressed the situation where the number of latent traits was four and only two factors were extracted. The second specialized design addressed the number of factors to be retained based on the eigenvalue-larger-than-1 criterion when the researcher has no previous knowledge of the number of underlying factors.

Four Latent Traits

Because in practice the number of latent traits can be higher than two, we analyzed a design for data with a four-dimensional latent trait structure. However, despite the four dimensions only two factors were extracted. We

assumed a 20-item questionnaire, in which the items had the same weights on the first two latent traits as with the Mix 3:1 case in the comprehensive design. Moreover, the first 5 items had weights B = 1, 2, 3, 4, 5 on the third latent trait, and the items 11 through 15 had weights B = 3, 6, 9, 12, 15 on the fourth latent trait. These choices imply that the items 1-5 are influenced by the four traits in the ratio 3:1:1:0; the items 6-10 in the ratio 3:1:0:0; the items 11-15 in the ratio 1:3:0:3; and the items 16-20 in the ratio 1:3:0:0.

The design factors were: (a) percentage of missingness, which was 5, 10, and 20, respectively; and (b) sample size, which was 50, 100, and 150, respectively. These sample sizes are different from sample sizes used previously because larger samples produced serious problems in the process of data generation, which required calculations based on large arrays needing too much memory capacity and/or computing time (using a SUN Unix system with 64 Mb memory). Listwise deletion was omitted from the analysis because, based on the results of the comprehensive design, we expected bad results here, and *EM* was omitted in the cell with 20 percent missingness and sample size 150 because computer memory problems were encountered while running the analyses.

The separation parameters for each item were fixed. Data were generated under a four-variate normal distribution with variances of 2.5 and covariances of 0.6 (also, see Equation 5). The relative expected frequency of nonresponse was fixed at REF-MAR (Table 4). Maximum likelihood factor analysis was used, followed by varimax rotation.

Eigenvalue Criterion

In practice, researchers often do not know the number of latent traits and thus rely on the eigenvalues of the factors for deciding on the number of factors to maintain for further analysis. To evaluate whether our comprehensive design would have led to other conclusions had we relied upon the eigenvalues rather than our a priori knowledge, a design was analyzed with factors: (a) 5 and 20 percent missingness; and (b) three different scoring weight configurations **B**, denoted Mix 1:0, Mix 3:1, and Mix 1:1 (see Table 3). It may be noted that 10 percent missingness was omitted; this was done to save computing time. Data were generated for 100 subjects, REF-MAR (Table 4), and all missing data methods except *LD*. The eigenvalues were evaluated for each covariance matrix corresponding to a particular missing data method.

Results

Results for the Comprehensive Design

Because differences between mean D^2s and between mean ϕs in corresponding cells of the designs for maximum likelihood factor analysis and principal component analysis often concerned the third decimal place, differences between results for maximum likelihood factor analysis and principal component analysis were considered ignorable; therefore, only maximum likelihood results are discussed. Also, the difference between D^2 results and between ϕ results based on varimax rotation and corresponding results based on procrustes rotation were ignorable. Consequently, only results for varimax rotation are discussed. The results for D^2 can be found in the Tables 5 and 6 (following pages). The main result was that in almost all design cells the *EM* algorithm had lower $\overline{D^2}$ and lower $s(D^2)$ than each of the other methods. This result should be kept in mind when reading the remainder of this section, where we mostly confine the discussion to imputation methods. This is done for each of the design factors: mixing configuration, percentage of missingness, sample size, and relative expected frequency of nonresponse.

Results for D^2

Mixing Configuration

For Mix 3:1 and Mix 1:0, *EM* had the lowest $\overline{D^2}$ and the lowest $s(D^2)$. For Mix 3:1, *PM* had the second lowest $\overline{D^2}$. The $s(D^2)$'s of *PM* and *IM* were the second lowest, and they were approximately equal. The $\overline{D^2}$ for methods *IM*, *CM*, *OM*, and *LD* roughly were two to three times as high as $\overline{D^2}$ for *PM*. For *CM* and *OM*, the $s(D^2)$ was twice as high as $s(D^2)$ for *PM*. *RI* always and *LD* often had the highest $\overline{D^2}$ and the highest $s(D^2)$. For Mix 1:0, *PM*, *IM*, and *CM* had approximately equal $\overline{D^2}$ and approximately equal $s(D^2)$. The $\overline{D^2}$ of *RI* and *LD* was much higher than $\overline{D^2}$ of *PM*, *IM* and *CM*. For Mix 1:1, *EM* and *PM* had the lowest $\overline{D^2}$ and the lowest $s(D^2)$. With a few exceptions, for *EM* and *PM* $\overline{D^2}$ and $s(D^2)$ were approximately equal. For the other methods, *IM*, *CM*, and *OM*, $\overline{D^2}$ often was at least ten times higher than for *PM*. For *RI* this factor was much higher. For the methods *IM*, *CM*, *OM*, and *RI*, $s(D^2)$ was approximately ten times higher than for *EM* and *PM*.

Table 5

$\overline{D^2}$ and $s(D^2)$ Across 50 Replications for Different Sample Sizes, Percentages of Missingness, and Mixing Configurations; Relative Expected Frequency Matrix MCAR. Entries are Result of Multiplication by 1000.

		Sample Size											
		100						500					
		Percent Missing											
		5		10		20		5		10		20	
Mix	Method	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$
1:0	*EM*	5	2	11	3	35	7	1	0	4	1	17	2
	PM	13	4	35	11	110	24	6	1	22	3	84	9
	IM	13	3	37	8	113	17	6	1	21	2	77	6
	CM	14	4	37	8	106	23	6	1	20	2	70	7
	OM	17	4	49	9	148	22	8	1	29	3	106	8
	RI	52	12	152	32	428	66	28	4	97	10	345	29
	LD	144	58					27	10	113	41		
	# omit	64	3	89	3	99	1	322	7	441	6	494	2

3:1	*EM*	2	1	4	1	11	3	0	0	1	0	5	1
	PM	10	4	23	8	65	18	4	1	15	2	49	6
	IM	14	4	41	8	142	19	7	1	25	2	101	6
	CM	18	6	47	13	142	35	9	1	29	4	104	11
	OM	25	7	67	13	227	33	13	2	45	4	170	10
	RI	75	23	215	45	664	84	43	5	152	12	535	37
	LD	67	35					10	5	40	18		
	# omit	64	3	89	3	99	1	321	8	439	6	494	2
1:1	*EM*	2	1	6	2	22	6	1	0	4	1	18	2
	PM	2	1	5	2	12	5	1	0	3	1	7	1
	IM	18	7	55	15	192	32	10	2	41	4	163	13
	CM	18	9	48	17	165	53	11	2	40	5	136	22
	OM	29	11	88	24	313	60	18	2	71	7	270	24
	RI	98	29	299	69	1012	178	71	9	252	22	901	65
	LD	33	19					6	3	27	13		
	# omit	65	4	89	3	99	1	321	6	440	6	494	2

Note: Rows "# omit" contain the average number of cases omitted and the accompanying standard error using *LD*. Empty cells could not be calculated due to singularity of the covariance matrix.

Table 6

$\overline{D^2}$ and $s(D^2)$ Across 50 Replications for Different Sample Sizes, Percentages of Missingness, and Mixing Configurations; Relative Expected Frequency Matrix MAR. Entries are Result of Multiplication by 1000.

		Sample Size											
		100						500					
		Percent Missing											
		5		10		20		5		10		20	
Mix	Method	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$
1:0	*EM*	5	2	12	3	33	7	1	0	5	1	17	2
	PM	15	4	36	10	106	23	6	1	22	3	83	11
	IM	15	5	37	8	113	21	6	1	20	2	75	7
	CM	16	5	36	8	102	23	6	1	20	3	67	7
	OM	21	8	50	9	153	25	9	1	30	4	107	9
	RI	55	16	151	31	449	65	29	5	99	10	341	23
	LD	126	44					23	7	78	34		
	# omit	59	3	80	3	95	2	292	8	400	8	475	5

3:1	*EM*	2	1	4	1	10	3	0	0	1	0	5	1
	PM	9	3	23	7	69	18	5	1	15	2	50	6
	IM	14	4	44	9	134	21	8	1	27	3	100	9
	CM	17	5	47	13	142	35	9	1	30	3	101	12
	OM	26	9	73	17	228	36	14	2	51	6	179	17
	RI	78	31	226	47	685	114	48	6	161	13	545	30
	LD	51	26					11	5	29	14		
	# omit	59	3	80	3	95	2	294	7	400	9	475	5
1:1	*EM*	2	1	7	2	22	6	1	0	5	1	18	2
	PM	3	1	8	3	18	6	1	0	4	1	13	3
	IM	22	7	64	17	237	61	14	3	50	7	195	19
	CM	19	7	58	16	167	53	12	3	43	8	146	21
	OM	36	11	105	23	358	81	23	4	83	12	310	27
	RI	107	30	321	66	1027	169	76	11	272	30	936	63
	LD	49	42					13	12	41	34		
	# omit	59	3	80	3	95	2	293	9	400	7	475	5

Note: Rows "# omit" contain the average number of cases omitted and the accompanying standard error using *LD*. Empty cells could not be calculated due to singularity of the covariance matrix.

The results confirmed the expectations as described in the section entitled "Implemented Missing Data Methods", except that for Mix 1:0, *PM*, *IM*, and *CM* had approximately equal $\overline{D^2}$ and approximately equal $s(D^2)$. If the correlation between the traits is higher than 0.24 used here, for Mix 1:0 *PM* may have the lowest $\overline{D^2}$ and $s(D^2)$ because then the traits are more alike. Table 7 contains the results of additional simulations where the correlation between the two traits was 0.5. For computational reasons, the sample size was kept at 100. The relative expected frequency of nonresponse was taken to be REF-MAR (Table 4), because this case is more realistic than MCAR. From Table 7, it can be concluded tentatively that, for higher correlation between the latent traits, *PM* had lower $\overline{D^2}$ and lower $s(D^2)$ than *IM* and *CM*.

Percentage of Missingness

For all methods, a higher percentage of missing item scores resulted in a higher $\overline{D^2}$ and a higher $s(D^2)$. For 10 percent missing, $\overline{D^2}$ was approximately three times higher than for 5 percent missing. For 20 percent missing, $\overline{D^2}$ was three to four times higher than for 10 percent missing. For 10 percent missing $s(D^2)$ was approximately two times higher than for 5 percent missing. With 20 percent missing, $s(D^2)$ was two to three times higher than for 10 percent missing. Thus, doubling the percentage of missingness at least doubled $\overline{D^2}$ and $s(D^2)$.

Sample Size

For sample size 500, both $\overline{D^2}$ and $s(D^2)$ were lower than for sample size 100 by a factor of approximately 0.5. These results probably are due to a reduction of chance capitalization with larger sample size. The conclusions with respect to the other design factors did not change.

Relative Expected Frequency of Nonresponse

Finally, for most methods $\overline{D^2}$ and $s(D^2)$ were approximately equal when relative expected frequency matrix MCAR (Table 5) was used to generate the missing item scores and when relative expected frequency matrix MAR (Table 6) was used. The exception was *LD*: the number of cases omitted under the MCAR assumption was higher than under the MAR assumption. For Mix 1:0 and Mix 3:1, this resulted in a lower $\overline{D^2}$ with MAR even though the estimators were biased.

Table7

$\overline{D^2}$ and $s(D^2)$ Across 50 Replications for Sample Size 100, Relative Expected Frequency Matrix MAR, and corr(θ_1,θ_2) = 0.5. Entries are Result of Multiplication by 1000.

		Percent Missing					
		5		10		20	
Mix	Method	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$	$\overline{D^2}$	$s(D^2)$
1:0	*EM*	5	2	12	4	35	8
	PM	11	3	30	8	88	24
	IM	15	4	37	8	115	21
	CM	14	4	38	8	103	23
	OM	19	5	52	10	153	25
	RI	54	15	138	25	433	63
	LD	128	65				
	# omit	59	3	80	3	95	2
3:1	*EM*	2	1	5	4	12	5
	PM	8	5	21	9	49	17
	IM	20	18	50	19	160	41
	CM	18	6	53	18	147	45
	OM	26	8	78	19	238	47
	RI	81	27	218	46	731	146
	LD	64	55				
	# omit	59	3	80	3	95	2
1:1	*EM*	2	1	6	2	18	6
	PM	2	1	6	2	14	6
	IM	21	8	64	18	228	49
	CM	21	9	57	18	177	50
	OM	33	12	102	29	346	64
	RI	102	30	317	63	1008	148
	LD	42	46				
	# omit	59	3	80	3	95	2

Note: Rows "# omit" contain the average number of cases omitted and the accompanying standard error using *LD*. Empty cells could not be calculated due to singularity of the covariance matrix.

Results for Tucker's ϕ

The mean value of Tucker's ϕ was calculated for each factor across 50 replications. In all cases, ϕ was higher than 0.98. ϕ was the closest to 1 for *EM* and it was the smallest for *LD*. However, the directions of the loadings matrices extracted from the complete data matrix and from the data matrices based on missing data methods were almost identical. Hence, application of a missing data method does not change the interpretation of a vector of loadings. Furthermore, the ϕs obtained for procrustes rotation and for varimax rotation were the same.

Results for Specialized Designs

Four Latent Traits

When the wrong number of factors was extracted, here two instead of four, *EM* no longer was the best method in terms of the lowest $\overline{D^2}$ and the lowest $s(D^2)$ (Table 8). Method *PM* consistently had the lowest $\overline{D^2}$. Method *CM* was the second-best method in 7 out of 9 cells and method *EM* was second-best in 2 cells. Method *RI* consistently had the highest $\overline{D^2}$. In most cells, Method *EM* had the lowest $s(D^2)$. Another method that performed consistently well was method *OM*. Keeping sample size constant, $\overline{D^2}$ and $s(D^2)$ both increased when the percentage of missingness increased; keeping percentage of missingness constant, $\overline{D^2}$ and $s(D^2)$ both decreased with increasing sample size. Compared with corresponding cells for $N = 100$ in the Comprehensive Design (Table 6), in Table 8 both $\overline{D^2}$ and $s(D^2)$ were considerably higher.

Thus, extracting the wrong number of factors may highly influence results for factor loadings matrices when *EM* or imputation methods are applied to incomplete data matrices. This could mean that the application of missing data methods as the ones studied here is most meaningful when the researcher already has an idea about the number of dimensions underlying his/her data.

Eigenvalue Criterion

The first five eigenvalues of the covariance matrices based on each of the five imputation methods *PM, IM, CM, OM,* and *RI* were calculated directly after the missing scores had been imputed. The *EM* method resulted in a completed data matrix after the algorithm had converged. Because the

Table 8

$\overline{D^2}$ and $s(D^2)$ Across 50 Replications of the Specialized Design Four Latent Traits. Entries are the Result of Multiplication by 1000.

	Sample Size																	
	50						100						150					
	Percent Missingness																	
	5		10		20		5		10		20		5		10		20	
Method	$\overline{D^2}$	s	$\overline{D^2}$	s	$\overline{D^2}$	s	$\overline{D^2}$	s	$\overline{D^2}$	s	$\overline{D^2}$	s	$\overline{D^2}$	s	$\overline{D^2}$	s	$\overline{D^2}$	s
EM	169	69	213	89	263	87	173	48	192	56	256	75	177	40	193	45		
PM	153	73	166	139	257	222	131	62	127	63	193	95	111	49	116	59	157	84
IM	202	73	236	94	361	101	207	70	218	72	283	66	185	49	209	48	256	58
CM	179	64	202	116	335	127	157	54	185	70	242	58	137	48	154	44	224	48
OM	186	62	217	109	389	96	160	53	197	68	293	65	141	46	167	46	273	52
RI	281	85	479	219	1117	291	221	72	351	94	829	182	188	50	319	55	770	145

researcher has to specify the number of factors in advance, we used the *EM* algorithm to obtain separate one-, two, three-, and four-factor solutions.

First, we discuss general results for imputation methods. For Mix 1:0, Table 9 shows that the first two mean eigenvalues were well over 1, and that the first mean eigenvalue was roughly two times the second mean eigenvalue. The third, the fourth, and the fifth mean eigenvalues were close to 1, and probably would be ignored by most researchers using the eigenvalue criterion. For Mix 3:1, the first mean eigenvalue often was very high, that is, in several cases it was higher than 15, whereas the second mean eigenvalue was approximately 2. The first mean eigenvalue was higher for Mix 3:1 than for Mix 1:0 because for Mix 3:1 all 20 items contributed to the first eigenvalue, whereas for Mix 1:0 only 10 items explicitly contributed to the first eigenvalue. For Mix 3:1, the third, the fourth, and the fifth mean eigenvalues were mostly lower than 1. Thus, they would be ignored in practical research. For Mix 1:1, in most cases the first eigenvalue was well over 1, whereas the other eigenvalues mostly were below 1. Again, researchers would draw the correct conclusion with respect to the number of factors to be retained.

For high percentage of missingness (20 percent), the eigenvalues were not that much different from eigenvalues obtained under low percentage of missingness (5 percent) so as to reach other conclusions concerning the number of factors to be retained. As for individual imputation methods, the differences between mean eigenvalues were not very impressive. Even method *RI* often led to eigenvalues that roughly reflect the correct number of factors. Finally, Table 10 shows that method *EM* led to results that reflect the correct number of factors underlying the data.

Discussion

In general, based on statistical considerations the *EM* algorithm has to be preferred over the other missing data methods studied for handling missing data in questionnaires when factor analysis of the data in envisaged. A version of the *EM* algorithm is available (see the Appendix) for estimating factor scores if questionnaire data are suffering from missing item scores. A drawback of the *EM* algorithm is that it converges slowly, and that convergence further slows down when the sample size and the percentage of missingness increase. With the ongoing increasing power and speed of computers, in the near future slowness is expected to become less of a problem. However, practitioners probably will continue having trouble understanding the algorithm and, moreover, because it is not readily available in statistical packages, use of the algorithm is problematic unless practitioners implement it themselves. Since

Table 9
Mean value Across 50 Replications of First Five Eigenvalues for Data Matrices Based on Imputed Scores.

		Percent Missingness									
		5					20				
		Eigenvalue number									
Mix	Method	1	2	3	4	5	1	2	3	4	5
1:0	*PM*	13.00	5.87	1.09	0.98	0.89	13.64	4.45	1.07	0.93	0.84
	IM	11.72	5.88	1.12	0.99	0.90	8.70	4.54	1.12	1.00	0.91
	CM	12.11	5.91	1.13	1.00	0.91	10.16	4.60	1.18	1.04	0.93
	OM	11.74	5.91	1.15	1.02	0.93	8.79	4.61	1.23	1.09	0.98
	RI	11.90	6.00	1.33	1.18	1.05	9.22	5.00	1.93	1.70	1.55
3:1	*PM*	17.29	2.17	0.57	0.36	0.31	17.70	1.58	0.66	0.44	0.38
	IM	15.58	2.20	0.60	0.39	0.34	11.32	1.70	0.66	0.55	0.48
	CM	16.10	2.21	0.62	0.41	0.36	13.25	1.73	0.78	0.58	0.51
	OM	15.59	2.22	0.66	0.44	0.38	11.43	1.78	0.89	0.67	0.58
	RI	15.65	2.34	0.80	0.60	0.53	11.80	2.19	1.46	1.24	1.10
1:1	*PM*	15.20	0.76	0.58	0.51	0.47	15.60	0.78	0.60	0.54	0.48
	IM	13.64	0.78	0.60	0.54	0.49	9.73	0.82	0.69	0.62	0.56
	CM	14.12	0.81	0.61	0.55	0.50	11.32	0.91	0.75	0.66	0.60
	OM	13.65	0.85	0.65	0.57	0.52	9.80	1.01	0.86	0.74	0.66
	RI	13.78	0.99	0.79	0.70	0.63	10.24	1.56	1.36	1.19	1.07

Table 10
Mean Value Across 50 Replications of First Five Eigenvalues for Data Matrices based on Score Imputation Using the *EM* algorithm.

		Percent Missingness									
		5					20				
		Eigenvalue number									
Mix	#f	1	2	3	4	5	1	2	3	4	5
1:0	1	12.81	6.11	1.07	0.94	0.85	13.09	4.23	0.96	0.85	0.78
	2	12.95	6.31	1.03	0.91	0.83	12.87	6.21	0.87	0.75	0.68
	3	13.18	6.38	1.09	0.92	0.82	12.77	6.12	1.10	0.79	0.71
	4	12.84	6.29	1.08	0.96	0.84	13.18	6.25	1.13	0.94	0.72
3:1	1	17.20	2.01	0.51	0.32	0.28	17.81	1.51	0.43	0.29	0.25
	2	17.93	2.14	0.54	0.31	0.27	17.61	2.30	0.39	0.26	0.22
	3	17.75	2.15	0.56	0.30	0.27	17.49	2.22	0.53	0.26	0.22
	4	17.76	2.19	0.58	0.32	0.27	17.87	2.18	0.57	0.32	0.22
1:1	1	15.86	0.72	0.54	0.48	0.44	15.44	0.57	0.46	0.41	0.37
	2	16.04	0.76	0.54	0.48	0.44	15.95	0.75	0.45	0.40	0.37
	3	15.61	0.74	0.55	0.48	0.43	15.49	0.75	0.55	0.41	0.36
	4	15.79	0.74	0.57	0.50	0.44	15.54	0.77	0.59	0.49	0.37

#f Displays the Number of Factors Retained

imputation is easier to understand and to implement, in practice a simple imputation method may thus be preferred.

The performance of the other methods mostly agreed with the expectations as expressed in the section entitled "Implemented Missing Data Methods". Method *PM* came out as the best imputation method, and thus is the method to be recommended for use in practical research when the researcher may not be in the position to use the superior *EM* algorithm. Method *LD* had the greatest variability in performance. It performed worst for Mix 1:0, where method *RI* even performed better. Also in agreement with the expectations was the result that, with the exception of method *LD*, missing data methods led to better results when the missing data mechanism was MCAR rather than MAR.

Other researchers (Brown, 1983; Finkbeiner, 1979; Lee, 1986; Muthén et al., 1987) also found that of the studied methods the maximum likelihood method, here implemented as an *EM* algorithm, performed best. It may be noted that these conclusions were not based on data generated under an IRT model and, moreover, with the exception of Finkbeiner's (1979) none of the other studies included imputation methods in the comparisons. Finkbeiner (1979) further concluded that method *IM* performed nearly as well as maximum likelihood estimation. We found that method *PM* was the second best method; this method was not studied by any of the other authors mentioned.

Our simulation study varied several factors believed to be important for comparing methods and held other factors constant, mostly because the design had to be manageable, but not because all fixed factors are unimportant. For example, we chose the latent trait distribution and the separation parameters (Equation 1) to be fixed. It can be argued, however, that in the unidimensional case increasing the variance of the latent trait relative to the separation parameters may lead to an increase of the variance of the item scores and an increase of the correlation between items. Increasing the spread of the separation parameters relative to the latent trait distribution will have the opposite effect. Since method *PM* uses the relation between the items and method *IM* does not, *PM* is expected to perform better when the variance of the latent trait increases and worse when the spread of the separation parameters increases, but *IM* will be unaffected. This example only shows the complexity of the problem under consideration, and also stresses the need for further research.

It was found that the results of the missing data methods with respect to $\overline{D^2}$ and $s(D^2)$ were the same for maximum likelihood factor analysis and principal component analysis. Moreover, varimax rotation of loadings matrices and procrustes rotation of a loadings matrix based on a missing data method to the complete data loadings matrix also led to the same results for

all methods with respect to $\overline{D^2}$ and $s(D^2)$. Thus, our conclusions are valid for principal component, factor analysis followed by varimax rotation, which is the factor analysis method most frequently used in psychological research.

When the researcher has no idea about the number of dimensions underlying the data, the extraction of the wrong number of factors on the basis of an imputed data matrix may lead to different results compared with results had the complete data been available. Fortunately, our analyses showed that the use of the popular eigenvalue-higher-than-1 criterion led to the correct conclusion about the number of factors to be retained. Since this result was obtained in the particular design studied here, however, the generalizability of the result to other situations is an open problem.

The present simulation study was based on missing data methods applied to multidimensional latent trait data. Cases of missingness were generated according to the MCAR and the MAR assumptions. Deviations from these assumptions introduce additional parameters, for example, describing the dependence of the probability of nonresponse on the item score and, as a consequence, additional decisions have to be incorporated in the design. This is left for future research (Bernaards & Sijtsma, 1999). The present study, however, revealed directions in which to turn when dealing with missing item scores from multidimensional questionnaires: (a) use the *EM* algorithm if possible; otherwise (b) impute the Person Mean.

References

Acock, A. C. (1997). Working with Missing Values. *Family Science Review, 10*, 76-102.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd Ed.). New York: Wiley.

Bello, A. L. (1993). Choosing among imputation techniques for incomplete multivariate data: a simulation study. *Communications in Statistics, 22*, 853-877.

Bentler, P. M. & Tanaka, J. S. (1983). Problems with EM for ML factor analysis. *Psychometrika, 48*, 247-253.

Bernaards, C. A. & Sijtsma, K. (1999). *Influence of simple imputation methods on factor analysis when item nonresponse in questionnaire data is nonignorable*. Submitted for publication.

Brown, C. H. (1983). Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika, 48*, 269-291.

Cattell, R. B. (1978). *The scientific use of factor analysis in the behavioral and life sciences*. New York: Plenum Press.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B, 39*, 1-38.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309-326.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika, 44*, 409-420.

Gleason, T. C. & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika, 40*, 229-252.

Greenless, J. S., Reece, W. S., & Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on variable being imputed. *Journal of the American Statistical Association, 77*, 251-261.

Huisman, M. (1998). Item nonresponse: Occurrence, causes, and imputation of missing answers to test items. Unpublished PhD-thesis, University of Groningen.

Huisman, M. & Molenaar, I. W. (1997). Imputation of missing data with item response theory models. Internal Report, University of Groningen.

Kelderman, H. & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika, 59*, 149-176.

Kim, J. & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research, 6*, 215-240.

Knol, D. L. & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*, 457-477.

Knol, D. L. & Ten Berge, J. M. F. (1989). Least-squares approximation of an improper correlation matrix by a proper one. *Psychometrika, 54*, 53-61.

Krzanowski, W. J. (1988). *Principles of multivariate analysis*. Oxford: Clarendon Press.

Lee, Sik-Yum (1986). Estimation for structural equation models with missing data. *Psychometrika, 51*, 93-99.

Likert, R. A. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 149-158.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Little, R. J. A. & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research, 18*, 292-326.

Little, R. J. A. & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (eds.), *Handbook of statistical modeling in the social and behavioral sciences* (pp. 39-75). New York: Plenum Press.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W.J. van der Linden & R.K. Hambleton (eds.), *Handbook of modern item response theory*. New York: Springer-Verlag.

Morrison, D. F. (1990). *Multivariate statistical methods*. Singapore: McGraw-Hill.

Muraki, E. & Carlson, J. E. (1995). Full-information factor analysis of polytomous item responses. *Applied Psychological Measurement, 19*, 73-90.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431-462.

Niesing, J. (1997). *Simultaneous component and factor analysis methods for two or more groups: a comparative study*. Unpublished doctoral thesis. University of Groningen.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika, 63*, 581-592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56*, 241-254.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*, 473-489.

Rubin, D. B. & Thayer, D. T. (1982). EM algorithms for factor analysis. *Psychometrika, 47*, 69-76.

Rubin, D. B. & Thayer, D. T. (1983). More on EM for ML factor analysis. *Psychometrika, 48*, 253-257.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.
Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393-408.
Tanner, M. A. (1996). *Tools for statistical inference. Methods for the exploration of posterior distributions and likelihood functions.* New York: Springer-Verlag.
Ten Berge, J. M. F. (1977). *Optimizing factorial invariance*. Unpublished doctoral thesis. University of Groningen.
Timm, N. H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. *Psychometrika, 35,* 417-437.
Tucker, L. R. (1951). A method for synthesis of factor analytic studies. *Personnel Research Section Report No. 984*. Washington, DC: Department of the Army.
Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics, 3,* 163-195.

Accepted October, 1998.

Appendix

The *EM* algorithm (Anderson, 1984; Bentler & Tanaka, 1983; Dempster, Laird & Rubin, 1977; Little & Rubin, 1987; Rubin & Thayer, 1982, 1983) always yields maximum likelihood estimates. *EM* treats the factor scores and the missing data as missing. Recall the classical exploratory factor analysis model,

$$\mathbf{X}_i = \mathbf{\Lambda}\mathbf{f}_i + \mathbf{U}, \quad i = 1, \ldots, N, \tag{7}$$

where $\mathbf{f}_i \sim N(0,\mathbf{I})$ and $\mathbf{U} \sim N(0,\mathbf{\Psi})$ and the $\mathbf{X}_i$ variables are centered around the mean $\boldsymbol{\mu}$. Under this model, assuming that the data and the factor scores $(\mathbf{x}_i,\mathbf{f}_i)$, ..., $(\mathbf{x}_N,\mathbf{f}_N)$ are jointly observed, the parameters to be estimated are $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. Note that given the factors, the variables are independent (conditional independence). Consequently, given $(\mathbf{\Lambda},\mathbf{\Psi})$ the loglikelihood to be maximized is equal to

$$\begin{aligned} L &= -N/2 \log|\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}| \; - 1/2 \sum_{i=1}^{N} \mathbf{x}_i' \left(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}\right)^{-1} \mathbf{x}_i \\ &= -N/2 \log|\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}| \; - 1/2 \,\mathrm{tr}\left[\mathrm{Cov}_{xx}\left(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}\right)^{-1}\right], \end{aligned} \tag{8}$$

where Cov_{xx} is the sample covariance matrix of the data. This likelihood is derived under the model and assumptions described here. It is not the likelihood of the MPLT model described in the section entitled "Generating the Data" which was used solely to generate the data.

The *EM* algorithm has two steps. First, in the *E* step the expected complete data sufficient statistics of the factor scores $\mathbf{f}_i$ and the data $\mathbf{X}_i$ are

calculated given the data $\mathbf{x}_i$ and the current estimates of the parameters $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. That is, the following expected matrices are calculated in the *E* step,

$$
\begin{aligned}
E\left(\mathrm{Cov}_{xx}|\mathbf{x}_1,...,\mathbf{x}_N,\mathbf{\Lambda},\mathbf{\Psi}\right) &= \mathrm{Cov}_{xx} \\
E\left(\mathrm{Cov}_{xf}|\mathbf{x}_1,...,\mathbf{x}_N,\mathbf{\Lambda},\mathbf{\Psi}\right) &= \mathrm{Cov}_{xx}\left(\mathbf{\Lambda\Lambda}'+\mathbf{\Psi}\right)^{-1}\mathbf{\Lambda} \\
E\left(\mathrm{Cov}_{ff}|\mathbf{x}_1,...,\mathbf{x}_N,\mathbf{\Lambda},\mathbf{\Psi}\right) &= \mathbf{\Lambda}'\left(\mathbf{\Lambda\Lambda}'+\mathbf{\Psi}\right)^{-1}\mathrm{Cov}_{xx}\left(\mathbf{\Lambda\Lambda}'+\mathbf{\Psi}\right)^{-1}\mathbf{\Lambda} \\
&\quad + \mathbf{I}-\mathbf{\Lambda}'\left(\mathbf{\Lambda\Lambda}'+\mathbf{\Psi}\right)^{-1}\mathbf{\Lambda}
\end{aligned}
$$

Second, in the *M* step the expected loglikelihood resulting from the *E* step is maximized just as if the factor scores were observed, that is, as if the expected covariance matrices from the *E* step were the observed ones. Using the distribution of the observation resulting from Equation 7, the estimators for $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ can be calculated from the conditional distribution of $\mathbf{X}_i$ given $\mathbf{f}_i$, using standard theory (Anderson, 1984, p. 37; Morrison, 1990, p. 92), resulting in

$$
\begin{aligned}
\mathbf{\Lambda} &= \mathrm{Cov}_{xf}\,\mathrm{Cov}_{ff}^{-1} \\
\mathbf{\Psi} &= \mathrm{Cov}_{xx}-\mathrm{Cov}_{xf}\mathrm{Cov}_{ff}^{-1}\,\mathrm{Cov}_{xf}'\,.
\end{aligned}
$$

Using the new value of the parameters, the missing data is estimated via

$$
\hat{\mathbf{x}}_i = \mathbf{\Lambda}'\left(\mathbf{I}+\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}\right)^{-1}\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{x}_i\,, \tag{9}
$$

the minimum variance estimator of the conditional expectation of **x** given **f**, $\mathbf{\Lambda}$, and $\mathbf{\Psi}$. The missing observations are replaced by their estimates from Equation 9, and we return to the *E* step. Iteration between *E* step and *M* step continues until convergence of the loglikelihood (Equation 8).